УДК 519.237.8+510.22

Д.А. Вятченин
Объединенный институт проблем информатики НАН Беларуси, г. Минск
viattchenin@mail.ru

Методология анализа данных, основанная на многоэтапной нечеткой кластеризации

В статье предлагается методология многоэтапного применения нечетких методов автоматической классификации в задачах интеллектуального анализа и обработки многомерных данных. Приводится результат вычислительного эксперимента при анализе искусственного набора данных и сформулированы предварительные выводы.

Введение

При решении различных социально-экономических задач, при проектировании разнообразных технических устройств и в процессе моделирования сложных систем особая роль отводится решению задач классификации, для решения которых традиционно применяются методы кластерного анализа, именуемые также методами распознавания образов с самообучением или методами автоматической классификации. Обрабатываемая информация зачастую оказывается неточной, нечеткой и противоречивой, что требует обращения к нечетким и возможностным методам автоматической классификации [1-4], в которых, в отличие от традиционных методов кластеризации, указывается степень принадлежности объекта кластеру, выражаемая, как правило, величиной из единичного отрезка вещественной прямой, что позволяет получить, с одной стороны, точные, а с другой – содержательно осмысленные результаты решения задачи классификации.

Вместе с тем кластеризация служит лишь средством решения задачи простой типологизации, то есть выявления стратификационной структуры исследуемой совокупности объектов, основанной на представлении классифицируемого множества в виде однородных групп объектов [5]. В таком случае решение задачи классификации является необходимым этапом исследования, предваряющим решение задачи структурной типологизации, то есть исследования структуры взаимосвязей полученных классов, включающего построение соответствующих иерархических систем – как на элементах классифицируемого множества, так и на классах элементов [5]. Таким образом, осуществление структурной типологизации множества объектов $X = \{x_1,\dots,x_n\}$ предполагает построение структурной классификационной схемы, которая определяется составляющими ее классами и взаимодействиями между классами – с одной стороны, а также объектами в пределах каждого класса – с другой.

Собственно задача структурной типологизации множества объектов не является новой – в [5] рассмотрены разнообразные варианты конечных прикладных целей для данной задачи классификации и изложен мощный статистический аппарат для ее решения. Однако в случае обращения для решения указанной задачи к методам нечеткой кластеризации представляется необходимым учитывать специфические особенности этих методов, связанные, в первую очередь, с интерпретацией результатов нечеткой кластеризации.

Целью данной работы является разработка общей схемы последовательного применения различных методов нечеткой и возможностной кластеризации в процессе анализа данных для решения задачи структурной типологизации исследуемой совокупности объектов.

Краткий обзор методов нечеткой кластеризации

Как и в традиционных методах кластерного анализа, в рамках нечеткого подхода к решению задачи автоматической классификации выделяются эвристическое, оптимизационное и иерархическое направления. Наиболее распространенным подходом к решению нечеткой модификации задачи автоматической классификации является оптимизационный подход, методы которого предусматривают нахождение оптимального, в смысле используемого критерия качества $Q(P(X))$, разбиения $P^*(X) = \{A^1, \ldots, A^c\}$ на заданное число c нечетких кластеров, описываемых функциями принадлежности μ_{li}, $l = 1,\ldots,c$, $i = 1,\ldots,n$, определенных на исследуемой совокупности объектов $X = \{x_1,\ldots,x_n\}$, так что задача нечеткой кластеризации заключается в нахождении экстремума целевой функции $Q(P(X))$, что в общем виде описывается формулой

$$Q(P(X)) \to \underset{P(X)\in\Pi}{extr}, \tag{1}$$

где Π – множество всех возможных нечетких разбиений $P(X)$ множества классифицируемых объектов X, при ограничениях, определяемых условием

$$\mu_{li} \geq 0, \ \sum_{i=1}^{n} \mu_{li} = 1, \ i = 1,\ldots,n, \ l = 1,\ldots,c, \tag{2}$$

именуемым также условием нечеткого c-разбиения или нечеткого разбиения в смысле Распини [3], которое описывается матрицей $P_{c\times n} = [\mu_{li}]$, где $\mu_{li} = \mu_{A^l}(x_i)$ – значение принадлежности элемента $x_i \in X$ некоторому нечеткому кластеру $A^l \in \{A^1, \ldots, A^c\}$.

При выборе вида функционала $Q(P(X))$ для проведения исследования учитывается, в первую очередь, вид матрицы исходных данных, а также вид шкалы, в которой измерены признаки объектов исследуемой совокупности. В силу ограниченности изложения дальнейшее рассмотрение предлагаемой методологии будет проводиться исходя из предположения, что исходные данные описываются матрицей «объект-признак», имеющей вид $\hat{X}_{n\times m} = [\hat{x}_i^t]$, $i = 1,\ldots,n$, $t = 1,\ldots,m$, так что каждый объект $x_i \in X$ может рассматриваться как точка в m-мерном признаковом пространстве $I^m(X)$. В случае, когда исходные данные представлены в форме матрицы «объект-объект» $\hat{\rho}_{n\times n} = [\hat{\rho}_{ij}]$, $i, j = 1,\ldots,n$, где общее обозначение $\hat{\rho}_{ij}$ используется вместо значений взаимных расстояний $\hat{d}_{ij}$ или коэффициентов сходства $\hat{r}_{ij}$ между объектами, общая схема анализа данных не претерпевает принципиальных изменений.

В случае, когда данные об исследуемой совокупности описываются матрицей вида «объект-признак», большинство критериев качества нечеткого разбиения в общем имеют вид

$$Q_*^{II}(P, \overline{\mathrm{T}}) = \sum_{l=1}^{c}\sum_{i=1}^{n} \mu_{li}^{\gamma} d(x_i, \overline{\tau}^l), \tag{3}$$

где x_i – элемент исследуемой совокупности, $\overline{\tau}^l$ – прототип нечеткого кластера $A^l \in P^*(X)$, и, как правило, в качестве $d(x_i, \overline{\tau}^l)$ используется квадрат какого-либо расс-

тояния. Как отмечает И.Д. Мандель, функционал (3) представляет собой «наиболее распространенный и изученный вариант экстремальной постановки задачи кластер-анализа в терминах размытых множеств» [6]. Некоторые модификации критерия (3) приведены в табл. 1.

Таблица 1 – Критерии качества нечеткого c-разбиения

Вид критерия	Параметры алгоритма	Ссылка
$Q^{II}_{FCM}(P,\overline{\mathrm{T}})=\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i-\overline{\tau}^l\right\|^2$	$2\le c<n$ – число классов; $1<\gamma<\infty$ – показатель нечеткости;	[1]
$Q^{II}_{BOFCM}(P,\overline{\mathrm{T}})=\zeta\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i-\overline{\tau}^l\right\|^2-$ $-(1-\zeta)\sum_{l=1}^{c}\sum_{a<l}\left\|\overline{\tau}^l-\overline{\tau}^a\right\|^2$	$2\le c<n$ – число классов; $1<\gamma<\infty$ – показатель нечеткости; $0<\zeta\le 1$ – параметр однородности;	[7]
$Q^{II}_{ICS}(P,\overline{\mathrm{T}})=\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\left(\mu_{li}^{\gamma}\left\|x_i-\overline{\tau}^l\right\|^2-\frac{\zeta}{c}\sum_{a=1}^{c}\left\|\overline{\tau}^l-\overline{\tau}^a\right\|^2\right)$	$2\le c<n$ – число классов; $1<\gamma<\infty$ – показатель нечеткости; $0\le\zeta$ – параметр классификации;	[8]
$Q^{II}_{PIM}(P,\overline{\mathrm{T}})=\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i-\overline{\tau}^l\right\|^2-$ $-\zeta\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}$	$2\le c<n$ – число классов; $1<\gamma<\infty$ – показатель нечеткости; $0\le\zeta$ – параметр классификации;	[9]

Одной из главных проблем при использовании оптимизационных методов является определение «реального» числа c нечетких кластеров, на которые «расслаивается» исследуемая совокупность, или, иными словами, проблема обоснования числа кластеров, встающая наиболее остро, когда исследователю число классов c вообще неизвестно. Для решения этой проблемы были предложены различные показатели, характеризующие получаемое при использовании того или иного алгоритма нечеткое разбиение $P^*(X)=\{A^1,\ldots,A^c\}$. В частности, при поиске нечеткого c-разбиения с помощью FCM-алгоритма, минимизирующего приведенный в табл. 1 критерий $Q^{II}_{FCM}(P,\overline{\mathrm{T}})$, а также его модификаций [1], [2], были предложены различные показатели, ряд которых приведен в табл. 2.

Таблица 2 – Показатели оптимальности числа классов в нечетком c-разбиении

Наименование показателя	Вид показателя	Решение задачи	Ссылка
Коэффициент разбиения	$V_{pc}(P)=\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{2}$	$\max_c\left(V_{pc}(P)\right)$	[10]
Показатель нечеткого гиперобъема	$V_{fh}(P)=\sum_{l=1}^{c}\sqrt{\det\left(\frac{\sum_{i=1}^{n}\mu_{li}^{\gamma}(x_i-\overline{\tau}^l)(x_i-\overline{\tau}^l)^T}{\sum_{i=1}^{n}\mu_{li}^{\gamma}}\right)}$	$\min_c\left(V_{fh}(P)\right)$	[11]

Продолж. табл. 2

Показатель толщины оболочки	$V_{st}(P)=\dfrac{\sum_{l=1}^{c}\left(\sum_{i=1}^{n}\mu_{li}^{\gamma}\left(\lVert x_i-\bar{\tau}^l\rVert-R_l\right)\Big/\sum_{i=1}^{n}\mu_{li}^{\gamma}\right)}{\sum_{l=1}^{c}\dfrac{R_l}{c}}$	$\min_c\left(V_{st}(P)\right)$	[12]
Показатель компактности и разделимости	$V_{cs}(P)=\dfrac{trace\left(\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}(\bar{\tau}^l-\bar{\tau})(\bar{\tau}^l-\bar{\tau})^T\right)}{\sum_{l=1}^{c}trace\left(\dfrac{\sum_{i=1}^{n}\mu_{li}^{\gamma}(x_i-\bar{\tau}^l)(x_i-\bar{\tau}^l)^T}{\sum_{i=1}^{n}\mu_{li}^{\gamma}}\right)}$	$\max_c\left(V_{cs}(P)\right)$	[13]

Если исходные данные представлены в форме матрицы «объект-объект» $\hat{\rho}_{n\times n}=[\hat{\rho}_{ij}]$, $i,j=1,\ldots,n$, то для решения задачи нечеткой кластеризации используются неметрические алгоритмы и соответствующие им показатели оптимальности числа классов, ряд которых рассматривается в работах [3], [4].

Касательно методов нечеткой кластеризации иерархического направления следует отметить, что соответствующие кластер-процедуры отличаются достаточно большим разнообразием – к примеру, различные иерархические кластер-процедуры основаны на различных определениях иерархии [14], [15], а относительно алгоритмов эвристического направления нечеткой кластеризации необходимо указать, что, как и в случае традиционного подхода к решению задачи автоматической классификации, эвристические методы нечеткой кластеризации играют большую роль на этапе разведочного анализа данных [5] – к примеру, MCM-алгоритм приближенной кластеризации [16] используется для определения числа классов c в искомом нечетком c-разбиении $P^*(X)$ и инициализации прототипов $\bar{\tau}^l$, $l=1,\ldots,c$ для последующей обработки данных оптимизационными методами нечеткой кластеризации, а D-AFC-TC-алгоритм возможностной кластеризации [17], [18] применяется для построения множества значений наиболее возможного числа нечетких кластеров A^l, $l=1,\ldots,c$ в искомом $P^*(X)$ [19].

Этапы структурной типологизации

Как указывалось выше, осуществление структурной типологизации предусматривает построение иерархий классов, а также иерархий объектов – элементов классов. В силу того, что иерархические кластер-процедуры обладают особенностью, заключающейся в резком возрастании, с ростом количества объектов классифицируемой совокупности, времени вычислений и требований к объему оперативной памяти ЭВМ, алгоритмы иерархического подхода применимы для классификации совокупностей объектов сравнительно небольшого объема и не могут быть прямо использованы для структурной типологизации больших массивов данных. Таким образом, проведение исследования с целью осуществления структурной типологизации множества объектов предполагает ряд этапов:

1) разбиение множества объектов на априори известное, или нет, число c классов;
2) построение иерархии на элементах каждого класса полученного разбиения;
3) построение иерархии классов полученного разбиения.

Указанные этапы структурной типологизации исследуемой совокупности объектов и последовательность их осуществления схематично изображены на рис. 1.

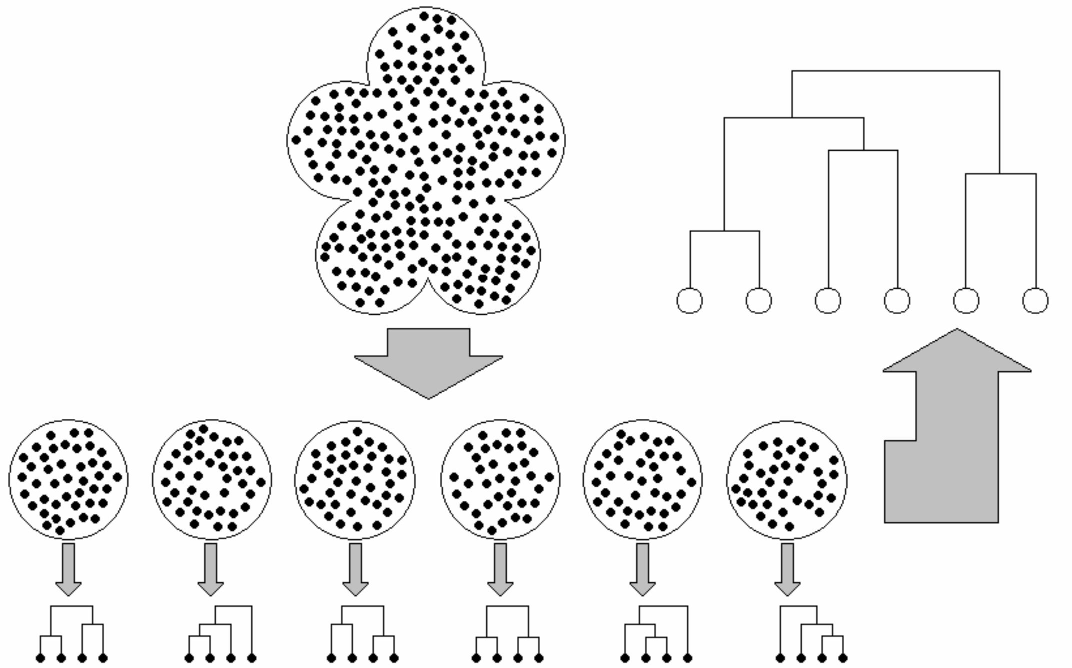


Рисунок 1 – Этапы осуществления структурной типологизации исследуемой совокупности объектов

Следует отметить, что в случае, когда число классов c априори неизвестно, то представляется необходимым проведение разведочного анализа данных с целью установления числа классов c, либо подмножества $C=\{c_*,\ldots,c^*\}$ возможных значений числа классов в искомом разбиении. Кроме того, если число классов оказывается сравнительно незначительным, а число объектов, попавших в каждый класс, достаточно велико, то этап разбиения на классы может быть повторно применен уже к множествам объектов – классам полученного разбиения; при этом возникает необходимость установления числа субкластеров в каждом классе и построение иерархии субкластеров в пределах каждого класса, что позволит проводить детальное исследование структуры классифицируемой совокупности.

В свою очередь, при построении иерархии классов в рамках структурной классификационной схемы, каждый класс может быть представлен либо его геометрическим центром, или, иными словами, прототипом, либо некоторым типичным для того или иного класса элементом, при этом нужно также отметить, что, помимо иерархических кластер-процедур, для выявления межкластерных связей могут применяться алгоритмы классификации на графах.

Необходимо, однако, указать, что в ситуации, когда искомая кластерная структура характеризуется таким видом неопределенности, как размытость [3], что требует обращения к методам нечеткой кластеризации, возникает проблема выделения значимых частей нечетких кластеров полученного в результате реализации первого этапа нечеткого c-разбиения $P^*(X)$ с целью проведения дальнейшего исследования и реализации второго и третьего этапов.

Концепция α-ядер нечетких кластеров

При обращении к оптимизационным методам нечеткой кластеризации, в силу условия нечеткого c-разбиения (2) каждому объекту $x_i \in X$ будет соответствовать вектор принадлежностей $(\mu_{1i},\ldots,\mu_{ci})$ классам полученного нечеткого c-разбиения $P^*(X)=\{A^1,\ldots,A^c\}$, вследствие чего возникает проблема интерпретации результатов классификации, то есть отнесения того или иного объекта к возможно меньшему числу классов. Наиболее распространенным методом интерпретации результатов является дефаззификация матрицы $P_{c\times n}=[\mu_{li}]$ нечеткого c-разбиения по правилу максимального значения принадлежности, выражаемого соотношением

$$P_i^{MM}=\mathrm{e}_l \Leftrightarrow \mu_{li}>\mu_{ai},\ a=1,2,\ldots,c,\ a\neq l, \tag{4}$$

так что значениями принадлежности μ_{li}^{MM} матрицы $P_{c\times n}^{MM}$ являются числа 0 и 1. Однако подобный подход является неприемлемым, если для некоторого объекта $x_i \in X$ его значения принадлежностей составляют $\mu_{li}=1/c$, $l=1,\ldots,c$. Кроме того, недостатком указанного подхода является утрата значений принадлежности объектов нечетким кластерам, позволяющая содержательно интерпретировать результаты кластеризации.

В свою очередь, концепция α-ядер нечетких кластеров, предложенная в [20], предполагает нахождение такого порога α, $\alpha\in(0,1]$, чтобы выполнялось условие

$$\sum_{l=1}^{c} card\Big(Supp(A^l(\alpha))\Big)\geq card(X), \tag{5}$$

где $X=\{x_1,\ldots,x_n\}$ – исследуемая совокупность объектов, α-ядра $A^l(\alpha)$, $l\in\{1,\ldots,c\}$ нечетких кластеров $A^l\in P$, $l\in\{1,\ldots,c\}$ для некоторого $\alpha\in(0,1]$ представляют собой нечеткие множества уровня, определяемые как $A^l(\alpha)=\{(x_i,\mu_{li}^{\alpha})\,|\,\mu_{li}^{\alpha}\geq\alpha\}$, так что $A^l(\alpha)\subseteq A^l$, $\alpha\in(0,1]$, $A^l\in\{A^1,\ldots,A^c\}$, и $Supp(A^l(\alpha))$ – носитель α-ядра $A^l(\alpha)$ нечеткого кластера $A^l\in P$, причем $Supp(A^l(\alpha))=A_{\alpha}^l$, то есть носитель α-ядра нечеткого кластера $A^l\in P$, $l=1,\ldots,c$ будет представлять собой α-срез $A_{\alpha}^l=\{x_i\in X\,|\,\mu_{li}\geq\alpha\}$ этого кластера при соответствующем значении α, а значения принадлежности объекта α-ядру определяются в соответствии с формулой

$$\mu_{li}^{\alpha}=\begin{cases}\mu_{li}, x_i\in A_{\alpha}^l\\ 0,\quad x_i\notin A_{\alpha}^l\end{cases}. \tag{6}$$

Порог α выбирается таким образом, чтобы каждый объект $x_i\in X$, $i=1,\ldots,n$ принадлежал бы по меньшей мере одному α-ядру нечеткого кластера, и может вычисляться по формуле

$$\hat{\alpha}=\min_i\max_l \mu_{li}, \tag{7}$$

что, в свою очередь, позволило сформулировать теорему [20], в соответствии с которой носители $\{A_{\alpha}^1,\ldots,A_{\alpha}^c\}$ α-ядер $\{A^1(\alpha),\ldots,A^c(\alpha)\}$ кластеров нечеткого c-разбиения

$P^*(X) = \{A^1, \ldots, A^c\}$ исследуемой совокупности объектов $X = \{x_1, \ldots, x_n\}$ образуют покрытие исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ в том и только в том случае, когда $\alpha \le \hat{\alpha}$, $\alpha \in (0,1]$, где $\hat{\alpha} \in (0,1]$ вычисляется по формуле (7).

Следствиями доказанной теоремы является ряд сформулированных также в [20] положений – в частности, если в условии (5) имеет место равенство, то носители $\{A_\alpha^1, \ldots, A_\alpha^c\}$ α-ядер кластеров нечеткого c-разбиения образуют разбиение исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ на непересекающиеся множества; кроме того, если $\alpha = \hat{\alpha}$, где значение $\hat{\alpha}$ вычисляется по формуле (7), то покрытие, образуемое носителями $\{A_\alpha^1, \ldots, A_\alpha^c\}$ α-ядер $\{A^1(\alpha), \ldots, A^c(\alpha)\}$ кластеров нечеткого c-разбиения $P(X) = \{A^1, \ldots, A^c\}$, минимально.

Концепция α-ядер нечетких кластеров позволяет, с одной стороны, отнести каждый объект исследуемой совокупности к наименьшему числу $\tilde{c}$, $1 \le \tilde{c} \le c$ нечетких кластеров нечеткого c-разбиения $P^*(X) = \{A^1, \ldots, A^c\}$, являющегося результатом классификации, а с другой – сохранить значения принадлежности μ_{li}, которые можно интерпретировать как степени обладания объектом $x_i \in X$ свойств класса, ассоциированного с нечетким кластером A^l, $l \in \{1, \ldots, c\}$ – элементом нечеткого c-разбиения $P^*(X)$.

Общая схема последовательного применения методов нечеткой кластеризации в процессе анализа данных

Как указывается в [20], «в случаях, когда объем исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ достаточно велик, носители $\{A_\alpha^1, \ldots, A_\alpha^c\}$ α-ядер $\{A^1(\alpha), \ldots, A^c(\alpha)\}$ кластеров нечеткого c-разбиения $P(X) = \{A^1, \ldots, A^c\}$ могут рассматриваться как множества объектов, подлежащие дальнейшей классификации». Данный тезис является отправной точкой при многоэтапном применении различных методов нечеткой кластеризации для осуществления структурной типологизации исследуемой совокупности объектов.

В случае отсутствия априорных предположений о числе классов в исследуемой совокупности объектов целесообразно построить подмножество $C = \{c_*, \ldots, c^*\}$ возможных значений числа классов в искомом нечетком c-разбиении, для чего можно воспользоваться предложенной в [19] методологией – в таком случае подмножество C будет представлять собой носитель нечеткого множества $\hat{V} = \{c_\ell, \mu_{\hat{V}}(c_\ell)\}$, $c_\ell \in C$, где значения функции принадлежности $\mu_{\hat{V}}(c_\ell)$ интерпретируются как степени адекватности значений числа $c_\ell \in C$ классов в искомом нечетком c-разбиении $P^*(X)$, что, с одной стороны, позволит содержательно оценить «степень возможности» того или иного значения числа классов $c_\ell \in C$, а с другой – применить (m)FCM-CV-алгоритм, позволяющий обрабатывать данные в полностью автоматическом режиме [19].

После разбиения исследуемой совокупности объектов с помощью некоторой кластер-процедуры группы оптимизационных методов нечеткой кластеризации на оптимальное, в смысле используемого показателя оптимальности, число c классов, сле-

дует выделить α -ядра нечетких кластеров полученного нечеткого c -разбиения $P^{*}(X)$, и дальнейший анализ проводить *отдельно для каждой группы объектов, являющихся элементами носителя соответствующего* α *-ядра*.

Следует, однако, отметить, что при необходимости построения нечеткого c -разбиения $P(X)=\{A^{1},\ldots,A^{c}\}$, в случае, когда носители α -ядер $\{A^{1}(\alpha),\ldots,A^{c}(\alpha)\}$ нечетких кластеров образуют покрытие классифицируемого множества объектов, то, при принадлежности объекта нескольким носителям α -ядер нечетких кластеров, отнесение объекта к тому или иному классу можно производить в соответствии с правилом максимальной принадлежности, которое в рассматриваемом случае можно сформулировать следующим образом: если некоторый объект $x_i \in X$, $i \in \{1,\ldots,n\}$ принадлежит носителю α -ядра более чем одного нечеткого кластера, то он должен быть отнесен к носителю α -ядра того нечеткого кластера, значение принадлежности μ_{li}^{α} которого в смысле определения (6) является наибольшим; при этом максимальное значение μ_{li}^{α} сохраняется, а значения принадлежностей α -ядрам других нечетких кластеров полагаются равными нулю; в случае же, если объект x_i принадлежит носителям α -ядер нескольких нечетких кластеров с одинаковыми значениями принадлежности μ_{li}^{α}, то x_i относится к каждому такому α -ядру с этим значением μ_{li}^{α}, и для таких элементов $x_i \in X$ строится матрица пересечений между классами. Сформулированное таким образом правило максимальной принадлежности позволяет сохранить значение μ_{li}^{α} элемента $x_i \in X$ для содержательной интерпретации результатов классификации. Подобным образом могут интерпретироваться результаты, полученные с помощью D-AFC(c)-алгоритма возможностной кластеризации [18], [21]. Если число объектов – элементов носителя α -ядра нечеткого кластера окажется приемлемым для использования иерархических кластер-процедур, то в подобной ситуации оказывается возможным построение иерархии на элементах каждого выделенного класса объектов; в противном случае каждый класс аналогичным способом разбивается на субкластеры, и дальнейший анализ производится для субкластеров.

Этап исследования межкластерных связей подразумевает замену каждого нечеткого кластера его прототипом и применением к соответствующим точкам пространства $I^{m}(X)$ некоторого иерархического алгоритма нечеткой кластеризации, либо процедуры классификации на нечетких графах. Подобный методологический прием, позволяющий существенно сократить число классифицируемых объектов с целью применения иерархических кластер-процедур, описан на примере решения задачи экономико-геологического районирования территории, изложенном в [6].

Общая схема осуществления структурной типологизации множества объектов с помощью методов нечеткой кластеризации представлена на рис. 2.

Необходимо указать, что, в силу большого разнообразия методов нечеткой кластеризации и существования нескольких видов кластерных структур, характеризуемых размытостью, конкретный вариант представленной схемы диктуется, во-первых, условиями решаемой задачи и целями исследования, а во-вторых – особенностями того или иного алгоритма, применяемого на каждом этапе. Следует также отметить, что этапы, выделенные пунктирными овалами, могут отсутствовать при проведении того или иного конкретного исследования; кроме того, каждый из указанных этапов может реализовываться с помощью различных кластер-процедур, что, в свою очередь, позволит углубить анализ и выявить устойчивость структуры, искомой на соответствующем этапе.

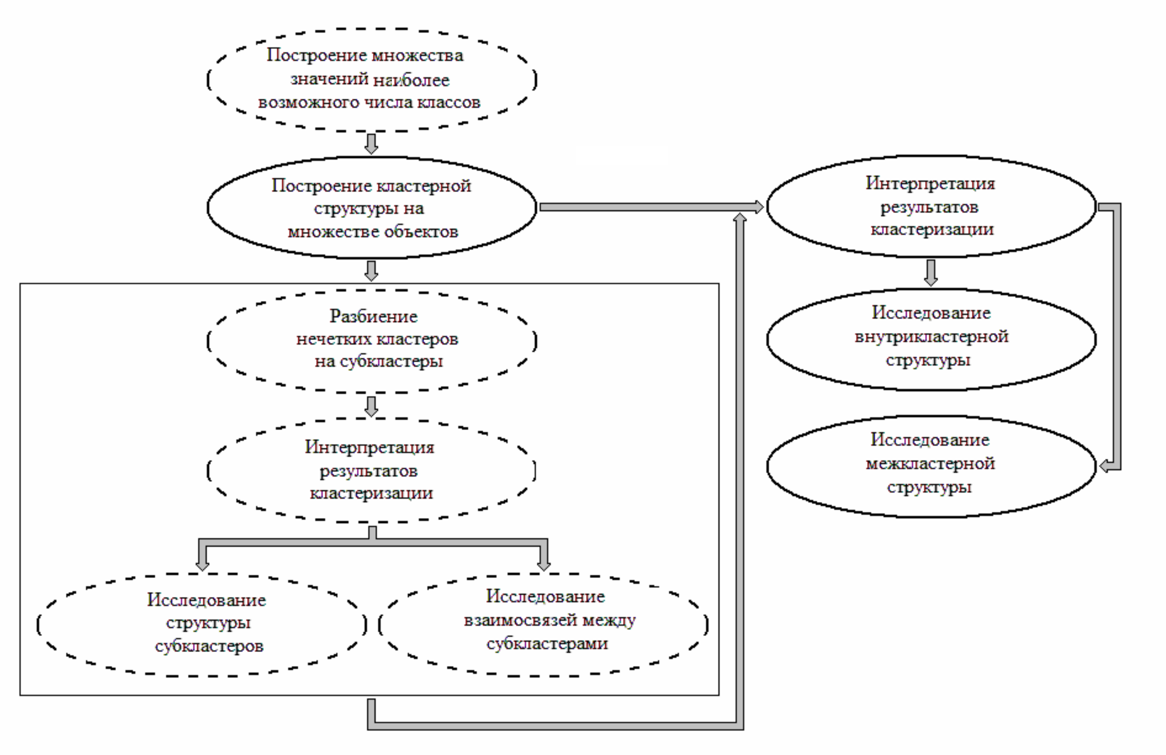


Рисунок 2 – Схема структурной типологизации исследуемой совокупности объектов, основанная на поэтапном применении методов нечеткой кластеризации

Иллюстративный пример

Представляется целесообразным проиллюстрировать сущность предложенной методологии на простом примере. Для проведения вычислительного эксперимента были выбраны изображенные на рис. 3 двумерные данные, представляющие собой совокупность 30 объектов $X = \{x_1, \ldots, x_{30}\}$.

Визуальный анализ данных демонстрирует, что число классов в искомом нечетком c-разбиении $P^*(X)$ может варьироваться от 2 до 6 – так, можно выделить два класса объектов, в первый из которых попадают объекты с номерами от 1 по 19, а во второй – с 20 по 30 включительно, с другой стороны, рассматривая разбиение множества X на три класса, можно выделить группы $\{x_1,\ldots,x_6\}$, $\{x_7,\ldots,x_{19}\}$ и $\{x_{20},\ldots,x_{30}\}$; в свою очередь, разбиение множества X на четыре класса состоит из групп $\{x_1,\ldots,x_6\}$, $\{x_7,\ldots,x_{11},x_{16},\ldots,x_{19}\}$, $\{x_{12},\ldots,x_{15}\}$, $\{x_{20},\ldots,x_{30}\}$, а разбиение X на пять классов – из групп $\{x_1,\ldots,x_6\}$, $\{x_7,\ldots,x_{11},x_{16},\ldots,x_{19}\}$, $\{x_{12},\ldots,x_{15}\}$, $\{x_{20},\ldots,x_{25}\}$, и $\{x_{26},\ldots,x_{30}\}$. При разбиении X на 6 классов выделяются скопления $\{x_1,\ldots,x_6\}$, $\{x_7,\ldots,x_{11}\}$, $\{x_{12},\ldots,x_{15}\}$, $\{x_{16},\ldots,x_{19}\}$, $\{x_{20},\ldots,x_{25}\}$, и $\{x_{26},\ldots,x_{30}\}$, однако объект x_8 занимает промежуточное положение между первой и второй группами, а объект x_{26} – между пятой и шестой, из чего явно следует размытость границ классов объектов, что предполагает необходимость обращения к нечетким методам классификации. Этапы осуществления структурной типологизации исследуемой совокупности, с указанием методов решения соответствующих задач, представлены в табл. 3.

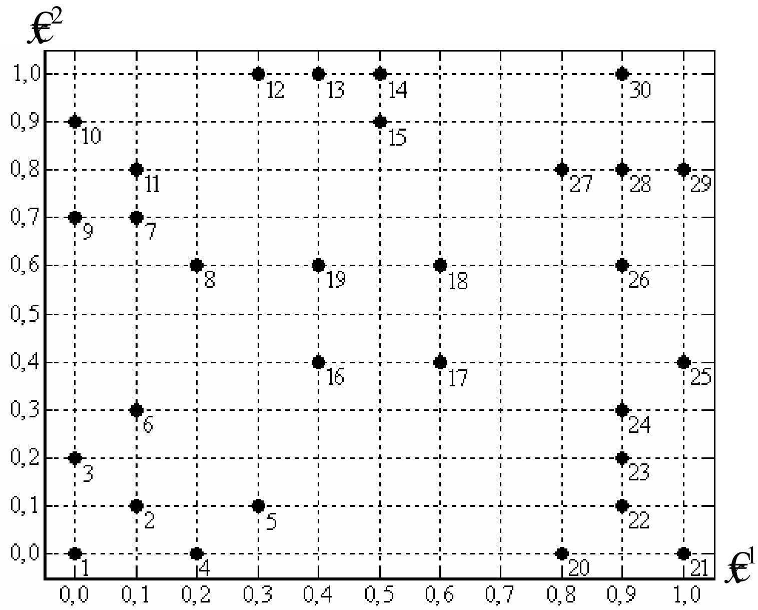


Рисунок 3 – Двумерные данные для вычислительного эксперимента

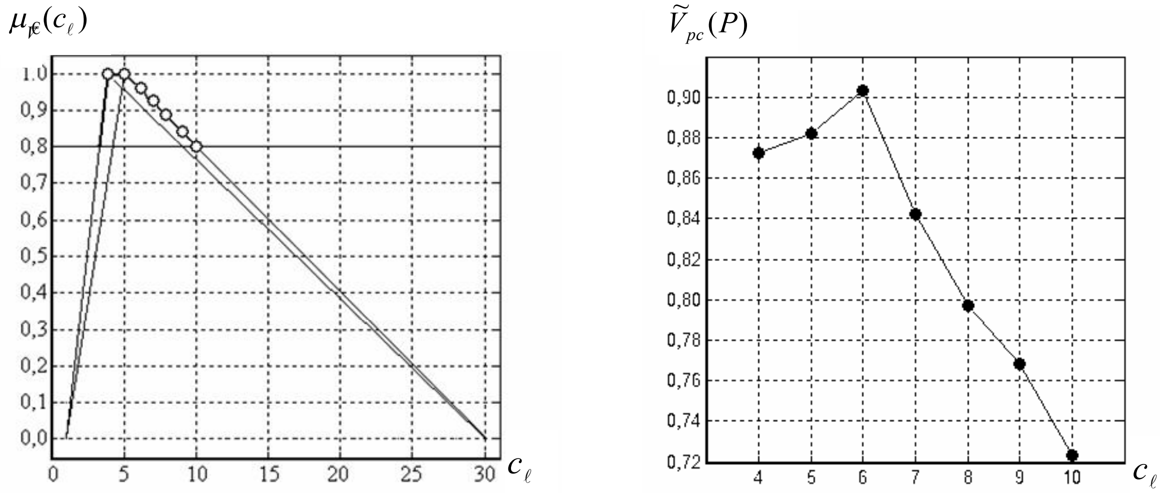
Таблица 3 – Этапы структурной типологизации анализируемой совокупности

Номер и содержание этапа		Метод решения задачи	Ссылка
1	Построение множества значений наиболее возможного числа классов	Построение нечеткого множества наиболее возможного числа нечетких кластеров с помощью D-AFC-TC-алгоритма возможностной кластеризации	[19]
2	Построение кластерной структуры на множестве объектов	Построение нечеткого c-разбиения на оптимальное число классов с помощью (m)FCM-CV-алгоритма нечеткой кластеризации	[19]
3	Интерпретация результатов нечеткой кластеризации	Выделение α-ядер нечетких кластеров	[20]
4	Исследование внутрикластерной структуры	Построение иерархий на подмножествах объектов – элементов носителей α-ядер нечетких кластеров с помощью H-AFC-TC-алгоритма возможностной кластеризации	[15]
5	Исследование межкластерной структуры	Построение иерархии на множестве прототипов нечетких кластеров с помощью H-AFC-TC-алгоритма возможностной кластеризации	[15]

При обращении к D-AFC-TC-алгоритму для построения нечеткого множества возможных значений числа классов $\hat{V}=\{c_\ell,\mu_{\hat{V}}(c_\ell)\}$, $c_\ell \in C=\{c_*,\ldots,c^*\}$, были проведены эксперименты с относительным обобщенным расстоянием Хемминга, относительным евклидовым расстоянием и относительной евклидовой нормой [19], [22]. При исполь-

зовании относительного обобщенного расстояния Хемминга было получено распределение $R^{*}(X)$ по $\hat{c}_1=5$ нечетким α-кластерам при $\alpha_1^{*}=0{,}8000$, а при использовании относительного евклидова расстояния и относительной евклидовой нормы было получено $\hat{c}_2=\hat{c}_3=4$ при $\alpha_2^{*}=0{,}8418$ и $\alpha_3^{*}=0{,}9750$ соответственно, что дало возможность построить нечеткие числа $V_1=(m_1=5,a_1=4,b_1=25)_T$ и $V_2=(m_2=4,a_2=3,b_2=26)_T$, объединение которых позволило построить нечеткую величину V с непрерывной функцией принадлежности $\mu_V(l)$, и далее – нечеткое множество $\hat{V}=\{c_\ell,\mu_{\hat{V}}(c_\ell)\}$, $c_\ell\in C=\{c_*=4,\ldots,c^*=10\}$ возможных значений числа классов. Следует указать, что носители нечетких α-кластеров – элементов соответствующих распределений $R^{*}(X)$ представляют собой элементы экспертных разбиений на 4 и 5 классов.

На рис. 4 а) символом ○ обозначены значения функции принадлежности $\mu_{\hat{V}}(c_\ell)$ нечеткого множества $\hat{V}=\{c_\ell,\mu_{\hat{V}}(c_\ell)\}$, $c_\ell\in\{4,\ldots,10\}$ возможных значений числа классов, а на рис. 4 б) – поведение для построенного $\hat{V}$ обобщенного коэффициента разбиения $\tilde{V}_{pc}(P)$ при обработке данных (m)FCM-CV-алгоритмом.



а) б)

Рисунок 4 – Построение нечеткого c-разбиения на оптимальное число классов:
а) нечеткое множество $\hat{V}=\{c_\ell,\mu_{\hat{V}}(c_\ell)\}$ возможных значений числа классов;
б) значения $\tilde{V}_{pc}(P)$ при обработке данных (m)FCM-CV-алгоритмом

В свою очередь, на рис. 5 изображены значения принадлежностей в смысле выражения (6) элементов исследуемой совокупности α-ядрам нечетких кластеров построенного с помощью (m)FCM-CV-алгоритма нечеткого c-разбиения $P^{*}(X)$. Значения принадлежности μ_{li}^{α} первого класса изображены символом ○, второго – символом ■, третьего – символом □, четвертого – символом ●, пятого – символом ∇, и, наконец, шестого класса – символом ▲. Пороговое значение, вычисленное по формуле (7), составило $\hat{\alpha}=0{,}79809$. Из рис. 5 очевидно, что носители α-ядер нечетких кластеров $A^l\in P^{*}(X)$, $l=1,\ldots,6$, не пересекаются и соответствуют группам объектов экспертного разбиения на 6 классов.

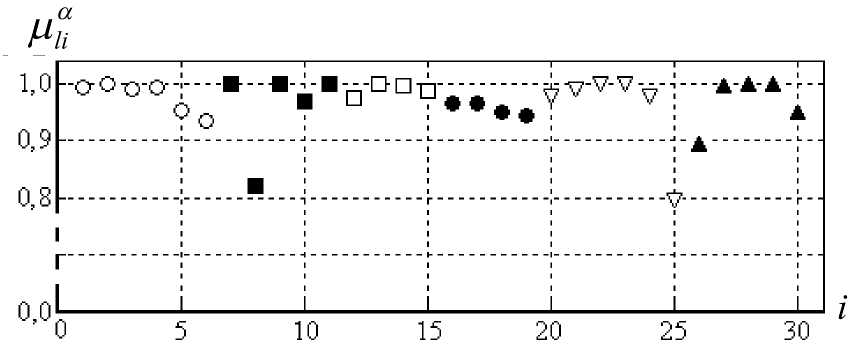
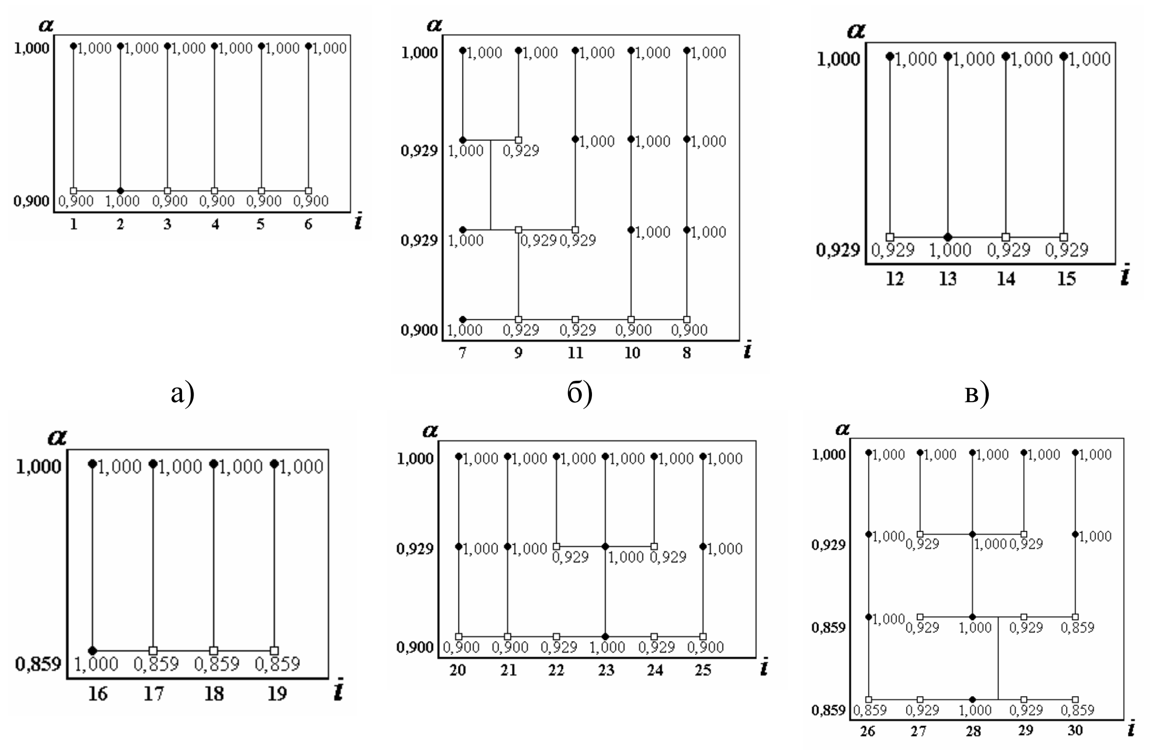


Рисунок 5 – Значения принадлежности объектов α -ядрам нечетких кластеров полученного нечеткого c -разбиения $P^*(X)$ на 6 классов

При обращении к H-AFC-TC-алгоритму возможностной кластеризации для выявления иерархической структуры классов объектов – носителей α -ядер нечетких кластеров $P^*(X)$ эксперимент проводился с использованием относительного евклидова расстояния. На рис. 6 а) – е) изображены иерархии распределений $R^*(X)$ по нечетким α -кластерам объектов, являющихся элементами носителей α -ядер нечетких кластеров – элементов построенного нечеткого c -разбиения $P^*(X)$. Типичные точки нечетких α -кластеров распределений $R^*(X)$ для каждого значения уровня сходства α обозначены символом ●, а объекты x_i, $i \in \{1,\ldots,30\}$ со значениями типичности, меньшими 1 – символом □.



а) б) в)

г) д) е)

Рисунок 6 – Иерархическая структура классов объектов – носителей α -ядер нечетких кластеров $P^*(X)$

Результаты классификации на данном этапе демонстрируют, что первый, третий и четвертый классы объектов однородны и обладают достаточно простой структурой, в то время как второй, пятый и шестой классы могут быть подразделены на субкластеры. Следует также отметить, что в построенных с помощью H-AFC-TC-алгоритма иерархиях содержательный смысл значений типичности объектов для нечетких α-кластеров распределений $R^*(X)$, как и смысл значений α, при которых эти распределения получены, отличны от значений принадлежности μ_{li}^{α} объектов α-ядрам нечетких кластеров и соответствующего значения порога $\hat{\alpha}$ [18].

Иерархическая структура прототипов $\bar{\tau}^l$, $l = 1,\ldots,6$ нечетких кластеров $A^1,\ldots,A^6$, координаты которых, вычисленные с помощью (m)FCM-CV-алгоритма, приведены в табл. 4, изображена на рис. 7. При проведении вычислительного эксперимента с H-AFC-TC-алгоритмом для построения иерархии прототипов $\bar{\tau}^l$, $l = 1,\ldots,6$ также было использовано относительное евклидово расстояние.

Таблица 4 – Координаты прототипов нечетких кластеров

Номер прототипа, l	Значения координат	
	$\hat{x}^1$	$\hat{x}^2$
1	0,114676	0,113194
2	0,072383	0,748145
3	0,426081	0,974915
4	0,499381	0,498146
5	0,911948	0,152280
6	0,900523	0,803509

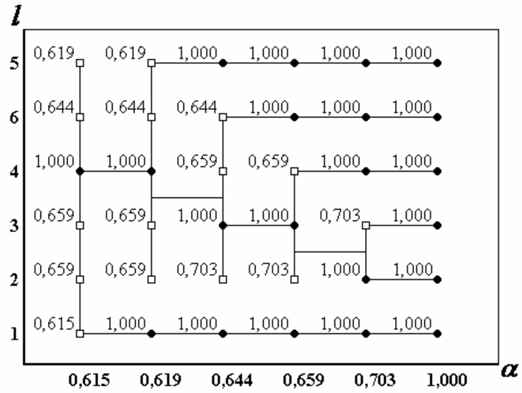


Рисунок 7 – Иерархическая структура прототипов нечетких кластеров

Представленная на рис. 7 иерархия прототипов отличается от иерархии рассмотренных выше экспертных разбиений объектов по классам, что очевидно уже при $\alpha = 0{,}619$, когда исследуемая совокупность прототипов расслаивается на два класса. Данное обстоятельство объясняется свойствами транзитивного замыкания, используемого H-AFC-TC-алгоритмом.

Заключение

Разнообразие методов нечеткой кластеризации позволяет использовать различные из них на каждом этапе исследования, и схема осуществления структурной типологизации, диктуемая целями исследования, характером данных и имеющейся априорной информацией, может варьироваться в каждом конкретном случае. Таким образом, предложенная методология многоэтапной нечеткой кластеризации может быть эффективно использована при решении самых разнообразных задач, таких, к примеру, как анализ данных в социально-экономических исследованиях, обработка результатов научных экспериментов, проектирование систем поддержки принятия решений, в том числе специального назначения, обработка и анализ изображений, а также для декомпозиции баз правил в системах нечеткого вывода. Помимо точности и релевантности результатов классификации, полученных на каждом этапе, главным достоинством предложенного подхода к анализу данных является возможность обработки данных в полностью автоматическом режиме, что открывает широкие возможности для решения задач в условиях реального времени.

Литература

1. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / Bezdek J.C. – New York : Plenum Press, 1981. – 230 p.
2. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition / F. Höppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester : Wiley Intersciences, 1999. – 289 p.
3. Вятченин Д.А. Нечеткие методы автоматической классификации / Вятченин Д.А. – Минск : УП Технопринт, 2004. – 219 с.
4. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J.M. Keller, R. Krishnapuram, N.R. Pal. – New York : Springer Science, 2005. – 776 p.
5. Прикладная статистика: Классификация и снижение размерности: [справ. изд]. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; под ред. С.А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с.
6. Мандель И.Д. Кластерный анализ / Мандель И.Д. – М. : Финансы и статистика, 1988. – 176 с.
7. Wang H.-F. Bi-criteria fuzzy C-means analysis / H.F. Wang, C. Wang, G.Y. Wu // Fuzzy Sets and Systems. – 1994. – Vol. 64. – P. 311-319.
8. Özdemir D. Fuzzy algorithms for combined quantization and dithering / D. Özdemir, L. Akarun // IEEE Transactions on Image Processing. – 2001. – Vol. 10. – P. 923-931.
9. Özdemir D. A fuzzy algorithm for color quantization of images / D. Özdemir, L. Akarun // Pattern Recognition. – 2002. – Vol. 35. – P. 1785-1791.
10. Dunn J.C. Well-separated clusters and the optimal fuzzy partitions / J.C. Dunn // Journal of Cybernetics. – 1974. – Vol. 4. – P. 95-104.
11. Gath I. Unsupervised optimal fuzzy clustering / I. Gath, A.B. Geva // IEEE Transactions on Pattern Analysis and Machines Intelligence. – 1989. – Vol. 11. – P. 773-780.
12. Davé R.N. Validating fuzzy partitions obtained through C-shells clustering / R.N. Davé // Pattern Recognition Letters. – 1996. – Vol. 17. – P. 613-623.
13. Bouguessa M. An objective approach to cluster validation / M. Bouguessa, S. Wang, H. Sun // Pattern Recognition Letters. – 2006. – Vol. 27. – P. 1419-1430.
14. Geva A.B. Hierarchical unsupervised fuzzy clustering / A.B. Geva // IEEE Transactions on Fuzzy Systems. – 1999. – Vol. 7. – P. 723-733.
15. Вятченин Д.А. Иерархический AFC-алгоритм нечеткой кластеризации, основанный на операции транзитивного замыкания / Д.А. Вятченин, П.Е. Савыгин, А.В. Шарамет // Сборник научных статей Военной академии Республики Беларусь. – 2006. – № 10. – С. 39-48.
16. Yager R.R. Approximate clustering via the mountain method / R.R. Yager, D.P. Filev// IEEE Transactions on Systems, Man, and Cybernetics. – 1994. – Vol. 24. – P. 1279-1284.
17. Вятченин Д.А. Прямые алгоритмы нечеткой кластеризации, основанные на операции транзитивного замыкания и их применение к обнаружению аномальных наблюдений / Д.А. Вятченин // Искусственный интеллект. – 2007. – № 3. – С. 205-216.
18. Вятченин Д.А. О возможностной интерпретации значений принадлежности в методе нечеткой кластеризации, основанном на понятии распределения / Д.А. Вятченин // Вести Института современных знаний. – 2008. – № 3. – С. 85-90.
19. Вятченин Д.А. Применение нечетких чисел для обоснования кластеров в методах нечеткой кластеризации / Д.А. Вятченин // Искусственный интеллект. – 2008. – № 3. – С. 523-533.
20. Вятченин Д.А. Метод мягкой интерпретации результатов нечеткой кластеризации / Д.А. Вятченин // Таврический вестник информатики и математики. – 2008. – № 1. – С. 107-114.
21. Viattchenin D.A. A new heuristic algorithm of fuzzy clustering / D.A. Viattchenin // Control & Cybernetics. – 2004. – Vol. 33. – P. 323-340.
22. Кофман А. Введение в теорию нечетких множеств / А. Кофман / [пер. с фр. ; под ред. С.И. Травкина]. – М. : Радио и связь,1982. – 432 с.

D.A. Viattchenin

Methodology of Data Analysis Based on Multistage Fuzzy Clustering

A methodology of automatic classification fuzzy methods multistage application in problems of intelligent analysis and processing of multidimensional data is proposed in the paper. The result of a numerical experiment for the analysis of the artificial data set is presented and preliminary conclusions are formulated.

Статья поступила в редакцию 01.06.2009.